PREDICT WHETHER A VARIANT WILL HAVE CONFLICTING CLINICAL CLASSIFICATION

INTRODUCTION :

In genetic diagnosis, accurate clinical judgment is dependent on the proper classification of germ-line variants. The germ-line variants are classified into three condensed categories i.e. benign or likely benign, Variant of Uncertain Significance(VUS), and likely pathogenic or pathogenic. In some cases, different labs can wrongly classify the germ-line variant for the same gene. For instance, if a doctor orders a genetic test from lab A for patient A and they find a genetic variant with "Pathogenic" classification. The doctor does a literature search and finds that this variant is likely to contribute to the patient's disease and decide to make some risk-reducing decision for patient A. Now, say a different doctor orders a genetic test from lab B for patient B and they find the same genetic variant,however, lab B classifies this variant as "Benign". This doctor might choose to seek further testing for the patient or do nothing at all. Such a classification is called conflicting clinical classification and it may cause flawed medical management.The research question of this project is to predict whether a variant will have a conflicting clinical classification.

My proposition is to apply machine learning methods i.e. classification technique in predictive analysis to identify the conflicting variant by comparing different lab results and thereby predicting whether a genetic variant will have a conflicting clinical classification in future analyses.

LITERATURE REVIEW:

Previous researches regarding germ-line variant classification have stated that there is discordance among clinicians and researchers in accurately classifying the variant. The difference in predicting the accurate germ-line variant by different labs have caused clinical impact on patients and among family members.

During a research regarding the inheritance risk of death due to Cardiomyopathy, a deceased patient's family were also tested for the presence of the pathogenic germ-line variant of the aforementioned condition. The family members who were negatively tested were considered not at risk and the positively tested ones were instructed to be regularly monitored for emergence of the disease. In due course of time, a different lab with the help of updated determination techniques has reclassified the pathogenic variant of a positively tested family member to 'likely benign' variant. A renewed test also determined that some previously negatively tested family members are now found to be at risk [*N Engl J Med, 2015*]. In another genetic variant detection study, its has been reported that concordance rates differ among clinical areas and variant types [*Genet Med, 2017*].

PREDICT WHETHER A VARIANT WILL HAVE CONFLICTING CLINICAL CLASSIFICATION

To overcome such problems different methods have been introduced based on guidelines such as interpreting germ-line variant based on American College of Medical Genetics and Genomics (ACMG) and Association of Molecular Pathology (AMP) guidelines [*Genet Med, 2017*], scoring system was introduced by labs based on ACMG guidelines [*Colleen Caleshu and Euan A. Ashley, 2016*] , 'in silico' algorithm was introduced by ACMG pathologists which is based on ACMG and AMP guidelines [*Genome Biology, 2017*].

Besides ACMG guidelines different laboratories have developed their own methods of variant classification, as variant-reporting guidelines did not address the weighting of evidence for variant classification. As an alternative, application of the automated tool for calculation of overall classification from evidence code was developed which is a pathogenicity calculator used to compare calculated ACMG-AMP classification based on the evidence code provided by the labs with final ACMG-AMP classification submitted by labs [*LM Amendola, 2016*].

Aiming to improve the interpretation process based on ACMG-AMP guidelines, web-based tools and software systems were developed. Prospective Registry of Multiplex Testing (PROMPT) is a web-based platform to assess the cancer risk of genetic variants [*J Clin Oncol, 2016*], ClinGen Pathogenicity calculator [*Genomic Medicine, 2017*], CharGer (Characterization of germ-line variant) etc are other software tools used for interpreting and predicting clinical pathogenicity of germ-line variant [*Adam D Scoff, Kuan Ling Huang, 2018*].

All these methods and software tools have helped to reduce discordance among different labs in predicting accurate germ-line variant to some extent. By applying machine- learning to predict a conflicting clinical variant will help the labs to accurately classify the germ-line variant.

DATASET :

The data-set that is used for this analysis can be found at Kaggle (https://www.kaggle.com/kevinarvai/clinvar-conflicting) and for further reference, raw ClinVar.vcf is used (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/clinvar.vcf.gz). Removed all variants from the original ClinVar.vcf which only had one submission as the problem only relates to variants with multiple classifications.

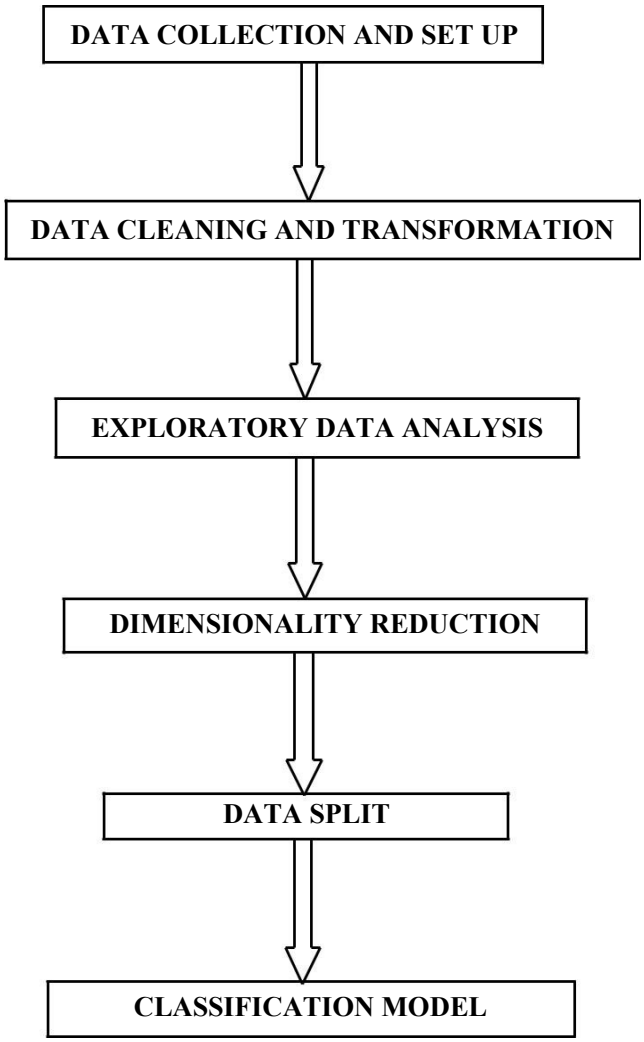
The data-set is comprised of 46 attributes and 65188 rows which consists of numerical as well as categorical values. It is represented as a binary classification problem. It is a binary representation of whether or not a variant has a conflicting clinical classification, here '0' represents consistent classifications and '1' represents conflicting classifications. These classifications has been assigned to the class column. The attributes that will be used for analysis is as follows: CHROM, POS, REF, ALT, AF_ESP Allele, AF_EXAC Allele, AF_TGP Allele, CLNVC, MC, ORIGIN,

PREDICT WHETHER A VARIANT WILL HAVE CONFLICTING CLINICAL CLASSIFICATION

CLASS, Allele, Consequence, IMPACT, SYMBOL, Feature_type, Feature Ensemble, BIOTYPE, EXON, INTRON, cDNA_position, CDS_position, Codons, Protein_position, Amino_acids, STRAND, PolyPhen, LoFtool, CADD_PHRED, CADD_RAW, SIFT, BLOSUM62, BAM_EDIT.

Descriptive statistics of the data-set and attributes is that there are missing values in the data-set. MOTIF_NAME, SSR, MOTIF_POS, HIGH_INF_POS, CLNDISDBINCL, CLNVI, MOTIF_SCORE_CHANGE, DISTANCE, CLNDISDB, CLNDNINCL has around 100 percentage missing values. The CLASS distribution is skewed a bit to the 0 class, which mean that there are fewer variants with conflicting submissions. As the mean of the target attribute (CLASS) is greater than that of the median and standard deviation is not equal to 1 it shows that our data-set is not normally distributed.

APPROACH :



PREDICT WHETHER A VARIANT WILL HAVE CONFLICTING CLINICAL CLASSIFICATION

STEP 1: DATA COLLECTION AND SET UP :

Data has been collected from Kaggle for analysis. Python platform has been set up to import and process the data-set. Libraries like Pandas, Numpy, Scipy have been imported to process the data-set for analysis.

STEP 2 : DATA CLEANING AND TRANSFORMATION :

As the data-set is real world data it contains missing values. The attributes with 100 percent missing value and few attributes with less significance has to be removed. Data transformation is done for converting data from one format to another format. Attributes that contain mixed data-types have to be converted to a single datatype, for instance, attribute that had integer as well as a string in it, has been converted to integer.

STEP 3: EXPLORATORY DATA ANALYSIS :

After cleaning and transformation of data, exploratory data analysis (EDA) has to be done. This includes the following methods :

- Univariate Analysis: This comprises of outliers analysis and treatment, Balancing dependent variables.
- Bivariate Analysis: This includes pair-wise relation among the variables, correlation analysis.
- Multivariate Analysis: This is used for analyzing relationships among more than two variables.

Other EDA methods such as normalization, graphical techniques such as box-plot,scatter-plot etc need to be performed.

STEP 4 : DIMENSIONALITY REDUCTION :

Dimensionality reduction is used to convert vast dimension of data into lesser dimensions without loss in information. The techniques used for dimensionality reduction are feature selection and feature extraction.

STEP 5 : DATA SPLIT :

Once dimensionality reduction is performed we need to split the data into training and testing set for further analysis. Training and test set must have similar predictors or variables.

STEP 6 : CLASSIFICATION MODEL :

After splitting the data, different classification model like decision tree, support vector machine has to be trained on the training set. The test set is used to

PREDICT WHETHER A VARIANT WILL HAVE CONFLICTING CLINICAL CLASSIFICATION

provide an unbiased evaluation of final model fit on the training set which will help to predict whether variant will have a conflicting clinical classification.